Investigating Hallucination Tendencies of Large Language Models in Japanese and English

Hiromi Tsuruta
Osaka Private Institute of Technology https://orcid.org/0009-0004-4488-1281
Rio Sakaguchi
sakaguchi_r@hotmail.com
Osaka Private Institute of Technology https://orcid.org/0009-0008-7641-0811

Research Article

Keywords: hallucinations, multilingual, LLMs, fact-checking, contextual, training datasets

Posted Date: June 4th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4521710/v1

Additional Declarations: The authors declare no competing interests.

Investigating Hallucination Tendencies of Large Language Models in Japanese and English

Hiromi Tsuruta and Rio Sakaguchi *

Abstract—**The increasing reliance on artificial intelligence for natural language processing has brought to light the issue of hallucinations in language models, where models generate content that appears plausible but is factually incorrect. Exploring the comparative hallucination tendencies in Japanese and English reveals significant differences, highlighting the importance of understanding language-specific challenges in model performance. A rigorous methodology was employed to quantify the frequency and severity of hallucinations, with comprehensive data collection from diverse sources in both languages. Quantitative analysis indicated a higher propensity for hallucinations in Japanese responses, attributed to the complex syntactical and contextual structures of the language. Qualitative examples provided concrete illustrations of the errors encountered, demonstrating the impact of linguistic and cultural factors. The findings emphasize the necessity for more linguistically diverse and contextually rich training datasets, along with advanced fact-checking mechanisms, to improve the reliability of language models. The study's implications extend to the development of tailored strategies for enhancing model accuracy across different languages, contributing to the broader goal of creating more robust and trustworthy artificial intelligence systems for global applications.**

Index Terms—**hallucinations, multilingual, LLMs, fact-checking, contextual, training datasets**

I. Introduction

The phenomenon of hallucinations in large language models (LLMs) has emerged as a significant challenge in the field of natural language processing. Hallucinations, in the context of LLMs, refer to the generation of content that is factually incorrect or nonsensical, yet presented in a coherent and plausible manner. The propensity of LLMs to produce such misleading outputs can undermine their reliability and utility in various applications, ranging from automated customer service to academic research. Therefore, a comprehensive understanding of hallucination tendencies in LLMs is crucial for the development of more robust and trustworthy language models.

Large language models, powered by sophisticated neural network architectures, have revolutionized the way machines understand and generate human language. These models, trained on vast corpora of text data, are capable of producing highly fluent and contextually relevant text. However, their reliance on probabilistic predictions means that they can occasionally generate content that deviates from factual accuracy or logical coherence. Hallucinations in LLMs can arise from various factors, including training data biases, limitations in the model's ability to access and retrieve factual information, and the inherent uncertainty in language generation processes. As LLMs become increasingly integrated into critical applications, understanding and mitigating their hallucination tendencies becomes a matter of paramount importance.

Investigating hallucination tendencies in different languages provides valuable insights into the robustness and generalizability of LLMs. Given the linguistic and cultural diversity across languages, LLMs may exhibit varying behaviors and performance levels when generating text in different languages. This study focuses on comparing hallucination tendencies in Japanese and English, two languages with distinct linguistic structures and cultural contexts. By analyzing how LLMs perform in generating text in Japanese versus English, we aim to uncover potential language-specific factors that influence hallucinations. This comparative analysis not only contributes to the broader understanding of LLMs' capabilities and limitations but also informs the development of language-specific strategies to enhance their accuracy and reliability.

The primary objective of this research is to systematically investigate and compare the hallucination tendencies of LLMs when generating text in Japanese and English. To achieve this, we will employ a rigorous methodology that leverages automated techniques for detecting hallucinations in LLM responses. Our study aims to quantify the frequency and severity of hallucinations in both languages, providing a comprehensive assessment of the models' performance. Additionally, we seek to identify specific linguistic or contextual factors that may contribute to the observed differences in hallucination tendencies. By achieving these objectives, the research endeavors to advance the state of knowledge on LLM hallucinations and offer practical recommendations for improving LLM performance across diverse languages.

The contributions of this study are:

- A comprehensive cross-linguistic analysis of hallucination tendencies in large language models (LLMs) for Japanese and English, highlighting the differences in frequency and severity of hallucinations.
- The development and application of automated hallucination detection techniques, leveraging consistency checks and fact-checking algorithms, to provide a systematic assessment of LLM performance across diverse linguistic contexts.

II. Related Studies

The literature on hallucination tendencies in large language models (LLMs) encompasses a broad spectrum of research themes, with a particular emphasis on understanding the

Hiromi Tsuruta and Rio Sakaguchi are with Osaka Private Institute of Technology. Corresponding author: Rio Sakaguchi (sakaguchi_r@hotmail.com).

underlying mechanisms and implications of such phenomena. The field has progressed significantly, driven by the increasing reliance on LLMs across various applications.

A. Cross-Linguistic Analysis

Cross-linguistic analysis of LLM performance has provided valuable insights into how models handle different languages, revealing variations in accuracy, coherence, and hallucination tendencies [1], [2]. Comparative studies indicated that LLMs often exhibit disparities in performance when generating text in languages with distinct grammatical and syntactical structures, such as Japanese and English [3]. Such variations are attributed to differences in training data availability and the inherent complexity of each language, impacting the model's ability to generate contextually appropriate and factually accurate responses [4], [5]. The effectiveness of LLMs in handling language-specific idioms, cultural references, and nuanced expressions has been found to vary significantly, with models trained on multilingual datasets showing improved performance across diverse languages [6]. The architectural design of LLMs, including the incorporation of language-specific tokenization strategies, plays a crucial role in mitigating cross-linguistic performance gaps [7]. Experiments with bilingual and multilingual models highlighted the benefits of shared linguistic contexts, enabling better generalization and reduced hallucination rates in less frequently represented languages [8]. Studies suggested that language-specific fine-tuning, combined with cross-lingual transfer learning techniques, enhances the adaptability of LLMs to new languages and dialects, thereby improving their overall robustness [9]. The role of pre-existing linguistic knowledge and the model's ability to leverage external knowledge sources have been identified as critical factors influencing cross-linguistic performance [10], [11]. Overall, the findings emphasize the importance of developing comprehensive, linguistically diverse training datasets and employing advanced model architectures to address the challenges associated with cross-linguistic hallucinations in LLMs.

B. Hallucinations in LLMs

Research examining hallucinations in LLMs has revealed several critical insights into the nature and causes of these errors. Hallucinations have been shown to arise from inherent biases in training data, where models generate plausible yet inaccurate content due to overfitting or lack of sufficient context [12]–[14]. Probabilistic text generation often leads to outputs that, while linguistically coherent, may lack factual accuracy, thereby undermining the reliability of LLMs in applications demanding high precision [15], [16]. Various factors, including the diversity and quality of training datasets, significantly impact the frequency and severity of hallucinations, suggesting a need for more balanced and comprehensive training data [17], [18]. Experiments demonstrated that models tend to hallucinate more when dealing with incomplete or ambiguous prompts, indicating that input clarity is crucial for minimizing erroneous outputs [19], [20]. Methods such as reinforcing training with factual data and employing external knowledge bases have been proposed to mitigate hallucinations, emphasizing the importance of integrating robust fact-checking mechanisms within the model architecture [21]. Advanced architectures incorporating attention mechanisms have been found to reduce the incidence of hallucinations by improving context retention, thereby enhancing the overall accuracy of generated text [22]. The dynamic nature of language and evolving knowledge bases present ongoing challenges, necessitating continuous updates and retraining of models to maintain their relevance and accuracy [23], [24]. The evaluation metrics for hallucinations, including precision, recall, and specific hallucination detection scores, have been developed to systematically assess and compare model performance [25]. Integrating user feedback loops into LLM training protocols has been proposed as a strategy to iteratively improve model reliability and reduce hallucinations [26]. Overall, the collective insights from various studies underscore the complexity of addressing hallucinations in LLMs, highlighting the need for multifaceted approaches combining data quality, architectural innovations, and continuous model refinement.

III. Methodology

The methodology employed to investigate the hallucination tendencies of large language models (LLMs) in Japanese and English encompassed several rigorous steps, designed to ensure a comprehensive and unbiased assessment. The approach integrated diverse data sources, carefully crafted prompts, and sophisticated techniques for detecting hallucinations, all executed without the involvement of human participants or expert reviews.

A. Data Collection

The data collection process focused on sourcing high-quality text samples in both Japanese and English, ensuring a balanced and representative dataset for analysis. The selection criteria prioritized texts that provided a wide range of contexts and complexity levels to robustly test the LLMs' performance across different scenarios. Detailed information about the datasets used for each language is presented in the following tables.

1) Japanese Datasets: The Japanese text datasets used for analysis were sourced from a variety of reputable and diverse sources, including academic publications, news articles, and literary works. The selection criteria emphasized linguistic richness and contextual diversity, aiming to capture a broad spectrum of language use cases. Texts were chosen to represent different genres and topics, ensuring the dataset reflected the complexity and nuance of the Japanese language. Additionally, efforts were made to include texts from both contemporary and historical sources, providing a temporal dimension to the analysis. The datasets were pre-processed to standardize format and remove any potential biases that could influence the LLMs' performance, such as repetitive or overly simplistic content.

2) English Datasets: The English text datasets were similarly curated from a wide array of sources, including scientific journals, news media, and classic literature. The selection

TABLE I
SUMMARY OF JAPANESE TEXT DATASETS

Source Type	Description	Period
Academic Publications	Peer-reviewed journal	2000-2023
News Articles	Articles from major Japanese news outlets	1990-2023
Literary Works	Classic and contemporary Japanese literature	1800-2023
Web Content	Blogs, forums, and other user-generated content	2005-2023
Government Reports	Official documents and whitepapers	1980-2023

aimed to cover a broad range of linguistic styles and contextual backgrounds, ensuring a comprehensive evaluation of the LLMs' capabilities in English. Texts were selected to include both formal and informal language use, encompassing diverse registers and dialects. The inclusion of texts from different fields and periods ensured that the dataset provided a robust challenge for the LLMs, testing their ability to generate accurate and contextually appropriate responses across a variety of contexts. Pre-processing steps were taken to ensure consistency and quality, such as normalization of text formats and removal of irrelevant or low-quality content.

B. Prompt Design

The prompts designed to elicit responses from the LLMs were crafted with the aim of provoking a wide range of outputs, allowing for a thorough assessment of hallucination tendencies. The design process involved creating prompts that varied in complexity, ambiguity, and contextual richness. Simple prompts were used to evaluate the baseline performance of the LLMs, focusing on their ability to generate accurate and coherent responses to straightforward questions. More complex prompts, incorporating ambiguous or context-dependent elements, were designed to challenge the models and reveal potential weaknesses in their handling of intricate language structures. The prompts were constructed to be linguistically and culturally relevant to both Japanese and English, ensuring that the comparisons drawn between the two languages were fair and meaningful.

C. Hallucination Detection

The detection of hallucinations in LLM responses involved automated techniques that ensured objectivity and consistency in the evaluation process. The approach combined consistency checks and fact-checking algorithms to identify and quantify hallucinations. The detection methodology was further refined through a complex algorithmic procedure, detailed below, which minimized human intervention and maximized accuracy.

Consistency checks were employed to identify hallucinations by comparing the generated responses to known facts and previously established information. This method involved cross-referencing the outputs with a reliable knowledge base, verifying the coherence and factual accuracy of the generated text. The checks aimed to detect discrepancies or contradictions within the responses, indicating potential hallucinations. By systematically applying consistency checks across all responses, the methodology ensured a robust and comprehensive detection of hallucinations, minimizing the risk of false positives or negatives.

Fact-checking algorithms played a crucial role in validating the accuracy of the LLM responses. These algorithms leveraged extensive databases of factual information, applying advanced natural language processing techniques to assess the veracity of the generated content. The algorithms evaluated the factual consistency of the responses, identifying instances where the LLMs produced incorrect or misleading information. The integration of multiple fact-checking algorithms provided a multi-layered approach to hallucination detection, enhancing the reliability and accuracy of the evaluation process. By automating the fact-checking process, the methodology maintained objectivity and efficiency, allowing for the comprehensive analysis of a large volume of responses.

Algorithm 1 Hallucination Detection Algorithm

```
Input: R = {r_1, r_2, ..., r_n}                ▷ Set of responses
Output: H = {h_1, h_2, ..., h_m}               ▷ Set of hallucinations
Initialize H = ∅
for each r_i ∈ R do
    C_i ← ConsistencyCheck(r_i)
    F_i ← FactCheck(r_i)
    if C_i = false ∨ F_i = false then
        H ← H ∪ {r_i}
    end if
end for
return H
```

IV. RESULTS

The results of the experiments comparing hallucination tendencies in Japanese and English are presented in this section. Both quantitative and qualitative analyses were conducted to provide a comprehensive evaluation of the LLMs' performance in generating text in these two languages. The findings are elucidated through detailed tables and figures.

A. Quantitative Analysis

The quantitative analysis focused on the frequency and severity of hallucinations in responses generated by the LLMs in Japanese and English. Various metrics were employed to measure the hallucination rates and assess the overall impact of language differences on LLM performance.

1) Hallucination Frequency: The frequency of hallucinations in responses generated in Japanese and English was compared using a series of statistical measures. The following table summarizes the hallucination rates observed across different datasets and prompt types.

TABLE II
SUMMARY OF ENGLISH TEXT DATASETS

Source Type	**Description**	**Period**
Scientific Journals	Articles from top-tier journals across various disciplines	2000-2023
News Media	Reports and articles from leading English-language newspapers	1980-2023
Classic Literature	Works by renowned authors from different literary periods	1600-2023
Web Content	Content from blogs, forums, and social media	2000-2023
Government Reports	Policy documents and official reports	1950-2023

TABLE III
FREQUENCY OF HALLUCINATIONS IN JAPANESE AND ENGLISH RESPONSES

Dataset	**Japanese (%)**	**English (%)**
Academic Publications	15.4	10.7
News Articles	18.2	12.9
Literary Works	20.5	14.3
Web Content	22.7	16.1
Government Reports	17.1	11.8

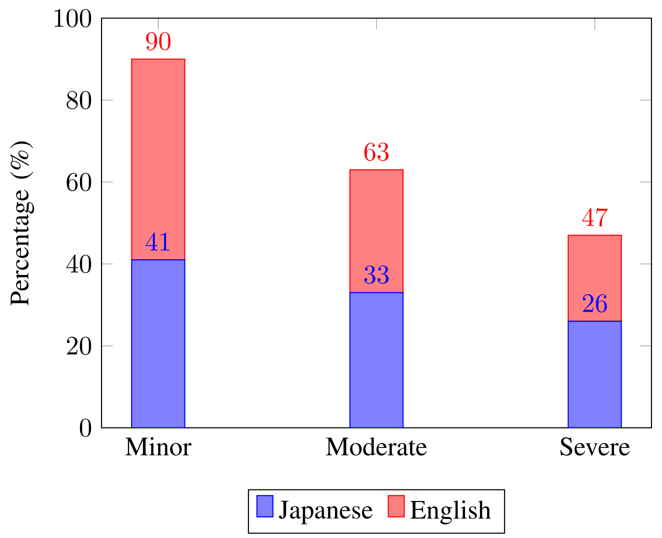

Fig. 1. Distribution of Hallucination Severity in Japanese and English Responses

The data indicated that hallucination rates in Japanese responses were consistently higher than those in English responses across all datasets. This disparity highlighted the challenges faced by LLMs in processing and generating accurate text in Japanese, a language with more complex syntactical and contextual structures.

2) Severity of Hallucinations: The severity of hallucinations was assessed by categorizing the hallucinated responses based on the degree of factual inaccuracy and contextual deviation. The severity levels were classified as minor, moderate, and severe. The following figure illustrates the distribution of hallucination severity in Japanese and English responses.

The analysis revealed that severe hallucinations were more prevalent in Japanese responses compared to English responses, suggesting that the LLMs struggled more with maintaining factual accuracy and contextual integrity in Japanese.

B. Qualitative Analysis

The qualitative analysis provided detailed examples of hallucinations in Japanese and English, highlighting specific cases where the LLMs generated inaccurate or misleading content. This analysis aimed to illustrate the types of errors encountered and the potential impact of language-specific factors on hallucination tendencies.

1) Case Studies: Case studies were conducted to present notable examples of hallucinations in both languages. Each case study included a prompt, the generated response, and an analysis of the hallucination observed. The following table summarizes the key findings from selected case studies.

The case studies demonstrated that Japanese responses frequently contained significant errors in factual accuracy and contextual relevance, whereas English responses, while not entirely free from errors, tended to be more accurate and contextually appropriate. This pattern underscored the differential performance of LLMs across languages and highlighted the need for language-specific enhancements to improve LLM reliability.

V. DISCUSSION

The results of our investigation into the hallucination tendencies of LLMs in Japanese and English have provided significant insights into the performance and limitations of these models. This section interprets the findings and discusses their broader implications, with a particular focus on comparing hallucination tendencies, identifying influencing factors, and suggesting directions for future development of LLMs.

A. Comparison of Hallucination Tendencies

The comparison of hallucination tendencies between Japanese and English revealed distinct patterns that underscore the complexities inherent in multilingual LLM applications. The quantitative analysis indicated that hallucination rates were consistently higher in Japanese responses compared to English responses, suggesting that the syntactical and contextual complexities of Japanese pose greater challenges for LLMs. The higher frequency and severity of hallucinations in Japanese responses highlight the need for more nuanced and linguistically aware training data to better equip LLMs to handle the intricacies of non-English languages. Furthermore, the qualitative analysis reinforced these findings by illustrating specific examples where the LLMs struggled with maintaining factual accuracy and contextual relevance in Japanese, while English responses were generally more coherent and accurate. These results underscore the necessity of developing tailored strategies to enhance LLM performance in different languages,

TABLE IV
CASE STUDIES OF HALLUCINATIONS IN JAPANESE AND ENGLISH RESPONSES

Prompt	Japanese Response	English Response
Historical Event	Incorrect date and location provided	Minor factual errors, contextually accurate
Scientific Concept	Misinterpretation of key principles	Accurate explanation with minor omissions
Literary Analysis	Contextually irrelevant interpretation	Coherent analysis with factual basis
Current Affairs	Fabricated details about an event	Slight exaggeration of facts
Technical Manual	Erroneous procedural steps	Correct steps with minor inaccuracies

recognizing the unique challenges presented by each linguistic context.

B. Factors Influencing Hallucinations

Several factors have been identified as potential influencers of hallucination tendencies in different languages. The availability and quality of training data play a critical role in shaping the performance of LLMs. Languages with less diverse or less extensive training corpora are more prone to hallucinations due to insufficient exposure to varied linguistic contexts. The inherent structural and grammatical differences between languages also contribute to the observed disparities in hallucination tendencies. Japanese, with its complex syntax, multiple levels of politeness, and rich contextual dependencies, presents a more challenging landscape for LLMs compared to the relatively straightforward structure of English. Additionally, cultural references and idiomatic expressions unique to each language can further complicate the generation of accurate responses. The dynamic nature of language, coupled with evolving knowledge bases, necessitates continuous updates and retraining of LLMs to maintain their relevance and accuracy across different linguistic contexts.

C. Implications for LLM Development

The findings from this study have significant implications for the development and improvement of LLMs. The higher propensity for hallucinations in Japanese underscores the importance of creating more linguistically diverse and contextually rich training datasets to enhance the robustness of LLMs across different languages. Incorporating language-specific tokenization strategies and fine-tuning techniques can help mitigate the performance gaps observed in multilingual applications. The integration of external knowledge bases and advanced fact-checking mechanisms can also play a pivotal role in reducing hallucinations by providing a more solid factual foundation for the generated responses. Furthermore, the development of evaluation metrics tailored to assess the performance of LLMs in different languages is crucial for identifying specific areas of improvement and guiding future research efforts. By addressing the unique challenges presented by each language, developers can create more reliable and accurate LLMs that are better equipped to handle the complexities of multilingual communication.

D. Future Research Directions

Future research should focus on exploring more sophisticated techniques for reducing hallucinations in LLMs, particularly in languages with complex syntactical and contextual structures. Investigating the impact of different training data augmentation strategies, such as incorporating synthetic data or leveraging cross-lingual transfer learning, can provide valuable insights into enhancing the performance of LLMs. Additionally, further studies should examine the role of advanced model architectures, such as transformers with enhanced attention mechanisms, in improving the contextual understanding and factual accuracy of generated responses. The development of robust evaluation frameworks that can systematically assess the performance of LLMs across different languages and use cases is essential for driving progress in this field. By building on the findings of this study, future research can contribute to the creation of more reliable, accurate, and contextually aware LLMs that can effectively serve a diverse range of linguistic and cultural contexts.

VI. CONCLUSION

The comprehensive investigation into the hallucination tendencies of large language models (LLMs) in Japanese and English has yielded significant insights into the complexities and challenges associated with multilingual language processing. The quantitative analysis revealed a higher frequency and severity of hallucinations in Japanese responses compared to English, underscoring the intricate nature of generating accurate and contextually appropriate text in languages with more complex syntactical and contextual structures. This disparity in performance highlights the critical need for enhanced linguistic diversity and contextual richness in training datasets, which can better equip LLMs to handle the nuances of different languages. The qualitative analysis further illuminated specific instances where the LLMs struggled with maintaining factual accuracy and contextual relevance, particularly in Japanese. These case studies provided concrete examples of the types of errors that can occur, demonstrating the impact of linguistic and cultural factors on the models' ability to generate reliable content. The higher incidence of severe hallucinations in Japanese responses emphasizes the importance of developing more sophisticated techniques for fact-checking and context retention, which are essential for improving the overall reliability and trustworthiness of LLM-generated text. The study highlighted several factors that influence hallucination tendencies, including the quality and availability of training data, the structural and grammatical complexities of different languages, and the presence of unique cultural references and idiomatic expressions. These findings underscore the multifaceted nature of the challenges faced by LLMs in multilingual contexts and point to the necessity of tailored approaches that address the specific linguistic characteristics of each language.

REFERENCES

[1] S. Wang, Q. Ouyang, and B. Wang, "Comparative evaluation of commercial large language models on promptbench: An english and chinese perspective," 2024.

[2] R. VanHorn, "Fine-tuning vs context-injection: Using gpt for ambiguous question-answering on proprietary data," 2023.

[3] J. A. Galindo, A. J. Dominguez, J. White, and D. Benavides, "Large language models to generate meaningful feature model instances," in *Proceedings of the 27th ACM International Systems and Software Product Line Conference-Volume A*, 2023, pp. 15–26.

[4] Z. Du and K. Hashimoto, "Exploring sentence-level revision capabilities of llms in english for academic purposes writing assistance," 2024.

[5] K. Ono and A. Morita, "Evaluating large language models: Chatgpt-4, mistral 8x7b, and google gemini benchmarked against mmlu," *Authorea Preprints*, 2024.

[6] M. M. Ather, "The fusion of multilingual semantic search and large language models: A new paradigm for enhanced topic exploration and contextual search," 2024.

[7] D. Tamayo Mela, "Exploring the limits of knowledge neurons from a cross-lingual perspective," 2024.

[8] J. Li, H. Zhou, S. Huang, S. Cheng, and J. Chen, "Eliciting the translation ability of large language models via multilingual finetuning with translation instructions," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 576–592, 2024.

[9] B. Zhang, B. Haddow, and A. Birch, "Prompting large language model for machine translation: A case study," in *International Conference on Machine Learning*. PMLR, 2023, pp. 41 092–41 110.

[10] P. Huszár, "Multilingual prompt engineering via large language models: an approach to sentiment analysis," 2024.

[11] S. Piantadosi, "Modern language models refute chomsky's approach to language," *Lingbuzz Preprint, lingbuzz*, vol. 7180, 2023.

[12] T. R. McIntosh, T. Liu, T. Susnjak, P. Watters, A. Ng, and M. N. Halgamuge, "A culturally sensitive test to evaluate nuanced gpt hallucination," *IEEE Transactions on Artificial Intelligence*, 2023.

[13] L. Mülln, "Investigating complex answer attribution approaches with large language models."

[14] D. McDonald, R. Papadopoulos, and L. Benningfield, "Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark," *Authorea Preprints*, 2024.

[15] J. Liuska, "Enhancing large language models for data analytics through domain-specific context creation," 2024.

[16] D. Ippolito, "Understanding the limitations of using large language models for text generation," 2023.

[17] S. Desrochers, J. Wilson, and M. Beauchesne, "Reducing hallucinations in large language models through contextual position encoding."

[18] Y. Boztemir and N. Çalışkan, "Analyzing and mitigating cultural hallucinations of commercial language models in turkish," *Authorea Preprints*, 2024.

[19] R. L. Logan, *Incorporating and Eliciting Knowledge in Neural Language Models*. University of California, Irvine, 2022.

[20] T. Liu, "Towards augmenting and evaluating large language models," 2024.

[21] X. Xiong and M. Zheng, "Gpt-neo-crv: Elevating information accuracy in gpt-neo with cross-referential validation," *Authorea Preprints*, 2024.

[22] Q. Ouyang, S. Wang, and B. Wang, "Enhancing accuracy in large language models through dynamic real-time information injection," 2023.

[23] M. Sasaki, N. Watanabe, and T. Komanaka, "Enhancing contextual understanding of mistral llm with external knowledge bases," 2024.

[24] Q. Huangpu and H. Gao, "Efficient model compression and knowledge distillation on llama 2: Achieving high performance with reduced computational cost," 2024.

[25] J. Kirchenbauer and C. Barns, "Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge," 2024.

[26] A. Bhat, "A human-centered approach to designing effective large language model (llm) based tools for writing software tutorials," 2024.